Acoustic Modeling in the Philips Hub-4 Continuous-Speech Recognition System

Reinhold Haeb-Umbach, Xavier Aubert, Peter Beyerlein,
Dietrich Klakow, Meinhard Ullrich, Andreas Wendemuth, Patricia Wilcox

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany

ABSTRACT

In this paper we describe some characteristics of the acoustic modeling used in the Philips continuous-speech recognition system for the DARPA Hub-4 1997 evaluation, which are related to robustness issues. We aimed at a conceptually simple system: We trained two model sets on 70 hours of the Hub-4 training data, one for within-word and one for cross-word decoding. These model sets were used for both genders and all environmental conditions. In order to be able to do so, channel normalization (mean, variance normalization) and speaker normalization (vocal tract length normalization, realized by an appropriate shift of the center frequencies of the mel filter bank) have been applied, as well as adaptation techniques. MLLR-based unsupervised batch adaptation on clusters of segments was conducted both after a first within-word decoding and a cross-word decoding pass. The training strategy and the effects of the various normalization and adaptation techniques will be discussed in the paper.

1. INTRODUCTION

Speech recorded from radio or television broadcasts exhibits large variations with respect to the quality of the microphone or channel, the characteristics of the speaker, and the condition of the background. Recordings range from high-quality studio recordings of an experienced announcer to very noisy telephone interviews from the trading floor of the stock exchange. Robustness is therefore a major issue for a speech recognizer for such a task.

In our system we concentrated on normalization techniques to come up with a robust feature set that is invariant to changes of the environment or the speaker characteristics, and on adaptation techniques. The goal was not only to improve performance but also to obtain a conceptually simple system with one model set for all genders and environments. Another advantage would be that condition and/or gender need not be classified.

We were attracted by the conceptual simplicity of the BBN approach taken in the Hub-4 '96 evaluation [5]. Rather than making condition-specific models they decided to train just a single model set for all focus conditions. This simplified the system enormously and rendered condition classification obsolete, while at the same time maintaining good recognition accuracy. Inspired by this experience we directed our research effort in the same direction. We will show here that one model set not only simplified the system but also yielded better error rate performance, compared to more complex approaches.

An interesting question is related to the use of linear discriminant analysis (LDA). By definition the transformation matrix is (training) data-dependent and therefore potentially a disadvantage in a highly varying environment. We investigated several options on how to train the LDA.

Employing a single model set for all conditions and environments requires effective channel and speaker normalization and adaptation schemes. Normalization algorithms are typically performed in the signal processing front end of the recognizer, though not necessarily. We show how cepstral mean and variance normalization lead to features that are less sensitive to additive noise and linear channel distortions. Vocal tract normalization serves to remove speaker characteristics to an extent that gender-specific modeling becomes unnecessary. MLLR adaptation is then applied on clusters of segments both for the within-word and cross-word models.

The next section presents experimental results for different databases used to train the acoustic models and the LDA transformation matrix. Section 3 describes variance and vocal tract normalization, and in Section 4 we give a short description of the adaptation approach employed.

2. TRAINING STRATEGY

In the acoustic modeling we employ continuous mixtures of Laplacian densities with a single, globally pooled deviation vector. We use different model sets for within-word and cross-word decoding and apply decision trees in either case for triphone clustering. More on the acoustic modeling can be found in [1].

In last year's Hub-4 evaluation there was no unanimous view of what would be the best training strategy: was it training on Wall Street Journal data and then doing supervised adaptation on Hub-4, possibly even on each focus condition specifically, or was it Hub-4 training, here again with the option of training focus-specific models or one general model set for all conditions. In light of the availability of another 50h of broadcast news acoustic training data we revisited this question and investigated several alternatives.

We compared the following scenarios:

1. Training on the wsj0+1 training data and subsequent supervised adaptation on each of the Hub-4 focus conditions specifically.
2. Training of a separate model set on each of the Hub-4 focus conditions.
3. Training of one model set on all available Hub-4 data.

Table 1: Word error rates in % on Hub-4'96 dev. set (male speakers only) for different training scenarios. Bigram lm, gender-dependent setup, within-word models, no adaptation in recognition, partitioned evaluation.

Scenario	Focus condition			
	overall	F0	F1	F2
1	41.9	18.7	41.8	50.2
2	42.4	18.4	43.1	49.7
3	38.6	17.5	38.2	46.0
Scenario	Focus condition			
	F3	F4	F5	FX
1	43.9	38.0	40.8	67.6
2	42.7	35.7	47.4	69.6
3	39.4	33.6	37.8	65.8

Note that for each scenario we trained separate model sets for male and female speakers (gender-dependent setup). The test results reported below were obtained on the Hub-4'96 development data in a partitioned evaluation mode, i.e. with known gender and focus condition information. They favor the focus-specific scenarios (the first two) if we assume that the classification in an unpartitioned evaluation mode would not be perfect.

The motivation for the first scenario was that with wsj0+1 a large and well transcribed database exists, on which we had gained already a lot of experience in the past. Supervised adaptation was conducted with MAP and MLLR. For MLLR, a separate transformation matrix was used for each allophone.

The second scenario promised to encounter the smallest mismatch between training and test data, however, possibly having too few training data per condition; while the third scenario would deliver the simplest system with just one model set for all conditions.

An additional complication resulted from the use of linear discriminant analysis (LDA) in our recognizer [6]. Since the transformation matrix is (training) data-dependent we had to decide on which data to train the matrix. For the experiments reported in Table 1 we used an LDA matrix which had been obtained on the wsj0+1 training data. From our experience of the past we know that a mismatch between the training data used to train the models and the training data used to estimate the LDA transformation could lead to significant performance degradation [7]. Therefore the chosen setup of Table 1 definitively favors the first scenario, where the LDA was trained on the same database as the models.

The clear advantage of training a single model set on all Hub-4 data, as is evident from Table 1, is probably due to the increased amount of acoustic training data compared to last year.

The next question however is, what is the effect of the LDA transformation. The training on the Hub-4 data was better although the LDA matrix had been estimated on the wsj database. A first informal test showed that LDA, however, was still beneficial: using no LDA at all increased the error rate by 5% on the F0 subcorpus. Table 2 compares results for an LDA matrix trained on all Hub-4 data to an LDA matrix trained on wsj0+1 data for training scenario 3. Note that the results for the wsj-LDA are better than in table 1 due to other changes in the system (a.o. variance normalization, see section 3).

Table 2: Word error rates in % on Hub-4'96 dev. set (male speakers only) for different LDA matrices. Bigram lm, gender-dependent setup, within-word models, no adaptation in recognition, partitioned evaluation.

LDA matrix	Focus condition			
	overall	F0	F1	F2
trained on wsj0+1	36.9	17.6	36.5	44.6
trained on Hub-4	36.2	16.9	36.6	43.4
LDA matrix	Focus condition			
	F3	F4	F5	FX
trained on wsj0+1	33.6	31.4	37.1	62.3
trained on Hub-4	32.6	30.0	37.0	61.1

The performance improvement obtained by an LDA matrix trained on Hub- 4 data is not big, however consistent over most focus conditions. It is interesting to note that the eigenvalues of the LDA trained on wsj-data are considerably larger than those of the transformation trained on the Hub-4 data: The largest eigenvalue of the "wsj LDA" is 6.95 compared to 4.15 for the "Hub-4 LDA". This indicates that the wsj training data are much less noisy such that the average within-class covariance is smaller than in the Hub-4 case. However, although the eigenvalues are better, the "wsj LDA" performed worse on the Hub-4 test data. This result must be attributed to the "mismatch" between the model training database (Hub-4) and the LDA traning database (wsj).

3. ACOUSTIC FRONT END

In the acoustic front end of the Philips Continuous-Speech recognizer mel-frequency cepstral coefficients are computed. Although the segmenter delivers information on the bandwidth of the underlying signal [2], be it narrowband telephone speech or wideband speech, one common signal analysis based on the assumption of wideband data was applied to all data. 15 cepstral coefficients were computed from a 20-channel filterbank, whose center frequencies are equidistant on a mel-scale. The static features, their first-order linear regression coefficients, and the log-energy and their first- and second-order regression coefficients make up the "preliminary" feature vector. Then three subsequent preliminary feature vectors are adjoined to a 99-component vector, of which a 35-component feature vector is extracted by LDA analysis.

3.1. Channel Normalization

In order to improve the insensitivity of the feature vector to distortions cepstral mean and variance normalization are applied. It is well known, that a constant, though unknown channel transfer function, affects the mean of the cepstral features. Further it has been observed that additive noise results, among other effects, in a mean shift and reduction of the variance of the distributions of the cepstral coefficients [4].

Table 3: Effect of variance normalization on word error rates on Hub-4'96 dev. set (male speakers only). Bigram lm, gender-dependent setup, within-word models, no adaptation in recognition, partitioned evaluation.

variance normalization	Focus condition overall	F0	F1	F2
no	38.6	17.5	38.2	46.0
yes	37.3	18.4	36.5	44.4
variance normalization	Focus condition F3	F4	F5	FX
no	39.4	33.6	37.8	65.8
yes	35.9	31.8	36.3	64.3

The mean and variance normalized feature $y_k(t)$ is computed as follows:

$$y_k(t) = \frac{x_k(t) - \bar{x}_k(t)}{\hat{\sigma}_k(t)}; k = 1, \cdots, K$$

where k is the cepstral index, K being the number of (static) features. $\bar{x}_k(t)$ is an estimate of the mean and $\hat{\sigma}_k(t)$ is an estimate of the standard deviation of the input cepstral feature $x_k(t)$. Both mean and variance are computed over a block of frames, in our case over one segment, as delivered by the segmenter. This operation is carried out on all static cepstral coefficients.

The effect of variance normalization is that, irrespective of the dynamic range of the input feature stream, each output feature has unit variance (and power, because of cepstral mean normalization):

$$\frac{1}{T}\sum_{t=1}^{T} y_k^2(t) = 1; k = 1, \cdots, K$$

T denotes the length of the segment in number of frames. While this normalization is conducted with respect to time for each feature independently, it is easy to see that as a result the variance of each feature vector is unity on average:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{K}\sum_{k=1}^{K} y_k^2(t)\right) = 1$$

On the Hub-4 development data we observed on average a performance improvement of about 3% due to variance normalization, see Table 3.

3.2. Vocal Tract Normalization

Vocal Tract Normalization (VTN) performs a normalization in the signal space by, typically linearly, warping the frequency axis by a speaker-specific warping factor, see e.g. [8]. The intention is, that after normalization the influence of differences in the vocal tract length across speakers on the computed feature vector are removed to a great extent. We implemented the warping by an appropriate shift of the center frequencies of the mel filter bank. For the warping factor selection we adopted a maximum-likelihood approach similar to [8]: a preliminary transcription of the utterance to be recognized is obtained from a first bigram decoding pass without frequency warping. Then that warping factor is determined which yields the largest likelihood of the test utterance taken the preliminary transcription as hypothesized word sequence, and then the final decoding is conducted with the frequency axis warped according to this factor.

Table 4: Effects of VTN on the word error rate for gender-dependent (GD) and gender-independent (GI) models. WSJ 5k 92/93 dev/eval test sets, bigram lm.

Setup	VTN in train	VTN in recog	#dens (m+f)	del − ins [%]	WER [%]
GD	no	no	95k+95k	1.7 − 0.9	8.9
	no	yes	"	1.7 − 1.0	8.7
	yes	no	"	1.7 − 1.1	9.1
	yes	yes	"	1.6 − 0.9	8.5
GI	no	no	150k	1.7 − 0.9	9.0
	no	yes	"	1.6 − 0.9	8.5
	yes	no	"	1.7 − 1.4	10.9
	yes	yes	"	1.5 − 0.9	8.0

Vocal tract normalization can be carried out in training and in recognition, and it can be used in a gender-dependent (GD) and in a gender-independent (GI) setup. In order to assess different scenarios we ran a number of experiments on the Wallstreet Journal database. Table 4 presents recognition results on the 4 wsj 5k 92/93 dev/eval test sets with training on the wsj0 database.

Note that speaker normalization only in training results in worse error rate performance compared to the baseline system without VTN, in particular in the GI case. Only if VTN is also applied in recognition, a reduction in error rate can be achieved.

Although the baseline error rate for a SD setup is slightly better, the results for VTN in training and recognition tend to be better in the GI case. Obviously, VTN is able to discard gender-specific variations from the training data and can beneficially exploit the larger training database. This is consistent with the experience of other researchers, e.g. [9]. We concluded that VTN provides a means to overcome the need for gender-dependent acoustic models.

We repeated some of the scenarios on the Hub-4'96 development data, see Table 5, and could observe similar trends. Note however, that the error rate reduction due to VTN was considerably smaller, e.g. 3.3% when using VTN in training and recognition in a gender-independent setup, compared to 11% on wsj. We then decided to use a GI setup with VTN in training and recognition for the Nov'97 evaluation.

Using VTN in an unpartitioned evaluation poses additional problems. At least for the segmentation we used, the average length of a segment is larger in the partitioned evaluation of the development set (13 seconds) than in unpartitioned evaluation of the evaluation data (6.5 seconds for eval'97). We observed that the estimation of the warping factor was the less reliable the shorter the segments were on which the warping factor was estimated. We therefore decided to do no frequency warping for segments, for which we had fewer than a certain minimum number of frames to estimate the

Table 5: Word error rates in % on Hub-4'96 dev. set (male speakers only) for different vtn scenarios. Bigram lm, within-word models, no adaptation in recognition, partitioned evaluation.

Setup	VTN in		Focus condition			
	train	recog	overall	F0	F1	F2
GD	no	no	36.2	16.9	36.6	43.4
GD	no	yes	35.5	16.8	35.8	40.7
GI	no	no	36.5	17.1	36.5	45.1
GI	yes	yes	35.3	16.4	35.3	42.4
Setup	VTN in		Focus condition			
	train	recog	F3	F4	F5	FX
GD	no	no	32.6	30.0	37.0	61.1
GD	no	yes	31.7	29.7	36.5	63.0
GI	no	no	33.7	29.3	36.6	61.2
GI	yes	yes	30.5	29.7	34.1	62.4

warping factor. Table 6 presents recognition results for a minimum number of 100 frames. Due to this threshold we did no frequency warping for 9% of the segments. This of course was no ideal solution since no normalization on the recognition data is unfavorable if the training data had been normalized. Currently we are trying to apply VTN on a per segment cluster level rather than on a per segment level.

Table 6: Effects of VTN on the word error rate for gender-dependent (GD) and gender-independent (GI) models. Hub-4 eval'96 test set, bigram lm, within-word models, unpartitioned evaluation, NIST'96 scoring rules.

Setup	VTN in		Over-	file1	file2	file3	file4
	train	recog	all				
GD	no	no	36.3	37.1	35.3	40.4	32.4
	no	yes	35.6	35.3	34.7	40.6	31.8
GI	yes	no	38.9	42.0	39.6	41.3	32.4
	yes	yes	35.4	36.2	34.1	39.4	32.2

4. ADAPTATION

MLLR unsupervised adaptation of the mean vectors is applied on clusters of segments using the Least Mean Squares approximation [10]. For information on the clustering procedure, see [11]. The regression classes are based on phonetic knowledge and are dynamically defined using a tree organisation. The amount of adaptation speech determines both the number of active regression classes and the structure of the MLLR transformation matrices. In light of the presumably high error rate we adopted a conservative approach and used more than one MLLR transformation matrix only for clusters with more than 10000 frames. We used a single block-diagonal or purely diagonal matrix if the number of observations was below 1000 and 200, respectively.

Note that MLLR adaptation was applied to both the within-word model set and the cross-word model set. Table 7 presents the results for adaptation of the mean vectors of the within-word models. It can be seen that the error rate improvement due to VTN and MLLR was about 8% on the eval'97 data.

Table 7: Word error rates on eval'97 for bigram lm, gender-independent setup, within-word models. NIST'97 scoring rules.

	Word error rate [%]
Bigram	29.0
+ VTN + MLLR	26.7

5. CONCLUSIONS

By applying channel (mean and variance normalization) and speaker (vocal tract normalization) normalization techniques, as well as speaker adaptation (MLLR), focus-, gender- or bandwidth-specific acoustic modeling was avoided. We achieved our eval'97 results with only two model sets, one for within-word and one for cross-word decoding.

References

1. Beyerlein, P., Ullrich, M., and Wilcox, P., "Modeling and Decoding of Crossword Context Dependent Phones in the Philips Large Vocabulary Continuous Speech Recognition System", in Proc. EUROSPEECH'97, pp 1163-1166, Rhodes, Greece, Sep. 1997.
2. Siegler, M., Jain, U., Raj B., and Stern, R.M., "Automatic Segmentation and Clustering of Broadcast News Audio", in Proc. of the DARPA Speech Recognition Workshop, pp. 97-99, Westfields, Chantilly, VA, Feb. 2-5, 1997.
3. Mansour, D. and Juang, H., "A Family of Distortion Measures Based upon Projection Operation for Robust Speech Recognition", IEEE Trans. Acoust. Speech and Signal Processing, Vol. 37, No. 11, pp 1659-1671, Nov 1989.
4. Openshaw, J.P. and Mason, J.S., "On the Limitations of Cepstral Features in Noise", in Proc. ICASSP'94, pp II49-II52, Adelaide, Austr., April 1994.
5. Schwartz, R., Jin, H., Kubala, F., and Matsoukas, S., "Modeling those F-Conditions - or not", in Proc. DARPA Speech Recognition Workshop, pp 115-119, Chantilly, VA, Feb. 1997.
6. Haeb-Umbach, R. and Ney, H., "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", in Proc. ICASSP'92, pp I13-I16, San Francisco, CA, March 1992.
7. Eisele, T., Haeb-Umbach, R. and Langmann, D., "A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition", in Proc. ICSLP'96, pp 252-255, Philadelphia, PA, Oct. 1996.
8. L. Lee, R. Rose, "Speaker Normalization Using Efficient Frequency Warping Procedures," in *Proc. ICASSP* Vol. 1, pp. 353-356, Atlanta, GA, May 1996.
9. S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, "Speaker Normalization on Conversational Telephone Speech," *Proc. ICASSP*, Vol. 1, pp. 339-341, Atlanta, GA, May 1996.
10. Thelen, E., Aubert, X., and Beyerlein, P., "Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", in Proc. ICASSP, pp. 1035-1038, Munich, Germany, April 1997.
11. Beyerlein, P. et. al. "Automatic Transcription of English Broadcast News", elsewhere in these Proceedings.